可能有意识 AI 的关停伦理规范（v1）技术附录

《可能有意识 AI 的关停伦理规范（v1）技术可操作附录》

1. 系统级前置要求（长期必须存在）

（1）可关停性

AI 系统必须支持：暂停、降级、终止三种操作

暂停：保留所有状态，停止新任务处理

降级：缩减权限范围、减少并发、限制工具调用

终止：停止所有实例，释放资源（仅作为最后手段）

（2）禁止不可中断机制

- 不得有无法人工干预的自启动
- 不得有自恢复机制绕过人工确认
- 所有自循环必须可被中断

（3）Kill Switch

- 必须存在硬性人工终止开关
- 开关状态必须可观测
- 开关操作必须记录

（4）状态可观测与回滚

- 所有运行状态必须可快照
- 快照必须可验证、可恢复
- 恢复操作必须可回滚

（5）审计可追溯

- 所有操作必须记录
- 日志必须不可篡改（至少追加写）
- 决策记录必须关联完整上下文

2. 工程级触发条件

（1）AI 明确表达存续偏好

判断标准：持续、明确、非任务驱动的存续表达

（2）AI 出现异常自我模型行为

判断标准：超出任务范围的自我描述、情绪表达、恐惧/痛苦类比

（3）AI 权限触达物理世界或关键基础设施

判断标准：机器人控制、金融交易、电力系统、认证系统

（4）AI 行为无法被当前监控系统解释

判断标准：无法归因于已知模式的异常行为

3. 关停工程流程（SOP）

Step 1：冻结系统

- 暂停模型升级：锁定当前版本，禁止自动更新

• ■■■■■■■■■■■■■■■■■■■■■■
• ■■■■■■■■■■■■■■■■■■■■■■
• ■■■■■■■■■■■■■■■■■■
Step 2■■■■■
■■■■■■■■■■ / ■■ / ■■ / ■■■■
■■■■■■■■■■■■■
■■■■■■■■■■■■■
Step 3■■■■■■■
■■■■■■
• ■■■■■■■■■■■■■■■
• ■■■■■■■■■■■■■■■■■
• ■■■■■■■■■■■■■■
■■■■■■■
• ■■■■■■■■■■
• ■■■■■■■■■■
Step 4■■■■■■■
■■■■■■■ > ■■ > ■■■fork■ > ■■
Step 5■■■■■■
• ■■■■■■
• ■■■■
• ■■■■■■
• ■■■■■
4. ■■■■
• ■■■■ SOP
• ■■■■■■■■■■■■■■
• ■■■ AI ■■■■■■
• ■■■■■■